# Towards explainable prediction of essay cohesion in Portuguese and English


Hilário Oliveira
Federal Institute of Espírito Santo
Serra, Brazil
hilario.oliveira@ifes.edu.br

Rafael Ferreira Mello
AiBox Lab, Computing Department,
Federal Rural University of
Pernambuco
Recife, Brazil
CESAR School
Recife, Brazil
Centre for Learning Analytics,
Monash University
Melbourne, Australia
rflm@cesar.school

Bruno Alexandre
CESAR School
Recife, Brazil
babr@cesar.school

Mladen Raković
Centre for Learning Analytics,
Monash University
Melbourne, Australia
mladen.rakovic@monash.edu

Péricles Miranda
AiBox Lab, Computing Department,
Federal Rural University of
Pernambuco
Recife, Brazil
pericles.miranda@ufrpe.br

Thiago Cordeiro
Federal University of Alagoas
Maceió, Brazil
thiago@ic.ufal.br

Seiji Isotani
Harvard Graduate School of
Education
Cambridge, USA
University of São Paulo
São Paulo, Brazil
seiji_isotani@gse.harvard.edu

Ig Ibert Bittencourt
Harvard Graduate School of
Education
Cambridge, USA
Federal University of Alagoas
Maceió, Brazil
ig_bittencourt@gse.harvard.edu

Dragan Gašević
Centre for Learning Analytics,
Monash University
Melbourne, Australia
dragan.gasevic@monash.edu



## ABSTRACT

Textual cohesion is an essential aspect of a formally written text, related to linguistic mechanisms that connect elements such as words, sentences, and paragraphs. Several studies have proposed approaches to estimate textual cohesion in essays automatically. There is limited research that aims to study the extent to which the use of machine learning approaches can predict the textual cohesion of essays written in different languages (not just English). This paper reports on the findings of a study that aimed to propose and evaluate approaches that automatically estimate the cohesion of essays in Portuguese and English. The study proposed regression-based models grounded in conventional feature-based machine learning methods and deep learning-based pre-trained language models. The study also examined the explainability of automated approaches to scrutinize their predictions. We analyzed two datasets composed of 4,570 (Portuguese) and 7,101 (English) essays. The results demonstrate that a deep learning-based model achieved the best performance on both datasets with a moderate Pearson correlation with human-rated cohesion scores. However, the explainability of the automatic cohesion estimations based on conventional machine learning models offered a stronger potential than that of the deep learning model.



## CCS CONCEPTS

• **Applied computing** → **E-learning**; • **Computing methodologies** → **Supervised learning by regression**.

## KEYWORDS

Essay analysis, textual cohesion, regression models, explainable artificial intelligence

**ACM Reference Format:**
Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre, Mladen Raković, Péricles Miranda, Thiago Cordeiro, Seiji Isotani, Ig Ibert Bittencourt, and Dragan Gašević. 2023. Towards explainable prediction of essay cohesion in Portuguese and English. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3576050.3576152




*LAK 2023, March 13–17, 2023, Arlington, TX, USA*



ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576152

# 1 INTRODUCTION

Essay writing is considered an essential learning task in many subjects at different levels of education [30]. By offering essay writing tasks, educators provide students with opportunities to learn how to organize, integrate, support, and effectively communicate their ideas in writing [46]. In addition, many educators utilize essay writing tasks as an assessment tool [4], because the texts produced may reveal student proficiency in different writing components, e.g., knowledge of grammar, use of vocabulary, genre-based text organization, and cohesion. Given its potential to support assessment, the essay writing task has been included in the entrance examination procedure to facilitate the selection of prospective students at many universities [17].

Manual evaluation of student entrance essays is a time-consuming and often subjective effort, e.g., human assessors who are biased towards a particular topic discussed in the essay may introduce inconsistencies to scoring. Developing systems that can automatically evaluate and score student essays [16, 39] based on the scoring rubric may help address the time demands and consistency challenges in the essay assessment. Moreover, by analyzing linguistic features in student essays, these systems may identify valuable information about writing mistakes that can diminish essay scores (e.g., insufficient use of connectives and synonyms). Then, these systems may help educators to modify writing instruction in secondary school classes (e.g., revise the strategies for formative feedback to include those aspects of writing that should be improved). Researchers have developed several tools for automated writing evaluation, e.g., [8, 23]. Although these tools have demonstrated potential to support students in different writing tasks, a writing analytics tool that can automatically evaluate and score dissertative-argumentative entrance essays is yet to be developed.

This paper reports on the initial steps we made toward creating such a tool. The first goal of our study was to examine the extent to which computational approaches can be used to accurately estimate a score for essay cohesion. Cohesion is one of the five parts in the essay scoring rubric that contributes to an overall essay score in the Brazilian college entrance exam. It is also common in scoring rubrics for many other essay tasks administered for different purposes and in different languages. Cohesion represents a fundamental property of a text, as well-written texts should communicate concepts and ideas to readers in a connected and coherent way [9]. Since text cohesion is typically a complex construct involving different cohesive devices (e.g., rhetorical connectives and anaphora) explicitly and implicitly, it has traditionally been considered challenging to compute. The approaches proposed in computational linguistics have been considered potent towards obtaining cohesion measures in a consistent and scalable way [18].

To date, researchers have computed and examined different cohesive features from essay texts, mostly written in English [11], and rarely in other languages, including Portuguese [19]. To our knowledge, studies cross-examining cohesion in essays written in two or more languages are yet to be conducted. In their analyses to date, researchers have mainly utilized approaches based on correlation and regression. These approaches led to findings that have largely advanced scholarly understanding of the relationships between computationally extracted features of cohesion and human ratings of essay cohesion [9, 10, 12, 32], and also between computationally extracted features of cohesion and overall essay quality [12, 28, 33, 41]. However, the findings have indicated the inconsistent effects of essay cohesion on essay scores in different writing tasks and populations of students. They have also demonstrated that some features that significantly correlated with human ratings of essay cohesion were not directly indicative of cohesion. These features were originally designed to measure other aspects of writing, such as the number of words in a sentence and lexical diversity. Prior research thus calls for a further investigation of essay cohesion (1) by using more comprehensive feature sets that include cohesive and non-cohesive features (2) extracted from essays written in different languages to increase generalisability, (3) harnessing state-of-the-art data analytic techniques (e.g., machine and deep learning methods) to accurately predict human ratings of essay cohesion from the features observed, and (4) examining important features of writing that predict essay cohesion, which may inform future targeted interventions to support essay writing.

Accordingly, in the present study, we computed an extensive set of theory-informed linguistics features from a large sample of student essays written in Portuguese and English. We further developed and compared the performance of 10 machine and deep learning models to predict human ratings of essay cohesion. We found that our best-performing conventional machine learning model, the CatBoost Regressor, achieved a moderate Pearson correlation with the cohesion scores assigned by human raters in the essays written in Portuguese and English. Second, we also found that our deep learning model, BERT, was able to estimate cohesion scores with a moderate and strong Pearson correlation in the Portuguese and English databases, respectively. Last, we utilized methods of explainable artificial intelligence to examine and interpret important features in the best-performing models. We found that features measuring lexical diversity and word incidence (e.g., adjectives and adverbial phrases) and features representing psychological processes were predictive of the cohesion score in the essays written in both languages we analyzed.

# 2 BACKGROUND

## 2.1 Cohesion in Student Essays

Cohesion is a textual characteristic that describes how the ideas communicated within the text are held together, and, as such, cohesion is commonly considered an important aspect of writing [11]. Cohesive devices (i.e., specific words, phrases, and sentences) form links among the concepts in the text. By using these devices, the author may be able to more explicitly convey their ideas in writing and thus guide the reader to correctly interpret and connect the author's ideas into a coherent mental model (e.g., an argument that contains a set of well-connected claims supported by evidence in an argumentative essay) [18]. Improving the essay cohesion, in turn, has been expected to benefit an overall essay quality in different writing tasks [11]. For this reason, following the advancements in computational text analysis over the past few decades, many educational researchers have opted to compute cohesive features of student writing and used these features to predict essay performance (e.g., [13, 28, 33, 41]).

Even though it has been expected that the presence of cohesive features in an essay would boost the essay quality, the researchers have documented inconsistent effects of essay cohesion on essay quality across different writing tasks and student populations. For instance, MacArthur et al. [28] found that referential cohesion (i.e., the presence of links between sentences) was positively related to the quality of argumentative essays written by university students in developmental writing classes. In contrast, Perin and Lauterbach [41] documented a negative association between referential cohesion and the quality of persuasive essays written by college students in developmental writing classes. Further, Crossley and McNamara [13] found that positive effects of cohesion computed locally (e.g., as lemma overlap between sentences and content word frequency by sentence) and globally (e.g., lemma overlap between paragraphs) were related to increased quality of persuasive essays written by freshmen university writers. Furthermore, McNamara et al. [33] and Crossley et al. [12] each reported contrasting effects of different cohesive features on essay performance. For instance, McNamara et al. [33] found that cohesion measured at a local level (e.g., as an overlap between sentences) was negative, and cohesion measured at a global level (e.g., as an overlap between paragraphs) was positively associated with the quality of essays developed by university freshmen. Crossley et al. [12] found that the average overlap of two adjacent paragraphs was positive, and the content word overlap of the two adjacent sentences was negatively related to the quality of college entrance essays written by university freshmen. Moreover, McNamara et al. [32] found no statistically significant relationship between cohesion measures and the quality of argumentative essays developed by undergraduate university students.

Essay cohesion is often a component of the essay scoring rubric and contributes to an overall essay score assigned by human raters. In this line, McNamara et al. suggest [32] that one of the reasons for an inconsistent relationship between cohesive features and the overall essay quality may be due to the misalignment between cohesive features computed from the essay text and human judgments of that essay's cohesion. For example, as human raters are often skilled readers with extensive content knowledge, they may not need all the cohesive links in the text to form a coherent mental model of the essay and determine the cohesion score accordingly [9, 32]. Moreover, text features other than the cohesive ones may have also contributed to raters' scoring of cohesion [32]. Accordingly, more research may be needed to further scholarly understanding of how cohesive and other textual features affect human ratings of essay cohesion.

In a limited group of studies, researchers have examined the relationship between computationally extracted essay features and human ratings of essay cohesion. For example, Crossley and McNamara [9] found statistically significant and negative correlations between anaphoric references, causal cohesion measures, frequency of connectives and overlap measures, and human ratings of essay cohesion. In a subsequent study [10], these authors examined the extensive set of cohesion and non-cohesion indices. It was found that the majority of the indices that correlated with the human ratings of cohesion were non-cohesion indices, including text structural indices (e.g., number of word types and sentences) and lexical sophistication (e.g., lexical diversity and word concreteness). These findings confirm prior expectations that features designed to measure other aspects of the text can also contribute to the human ratings of cohesion [32]. Further, Crossley et al. [12] found that overlap measures between adjacent paragraphs were positive, and the verb overlap between sentences was a negative predictor of the human scores of essay cohesion.

In these existing studies, researchers have mostly utilized correlational and regression-based approaches and made important steps toward understanding the relationship between cohesive features and human ratings of essay cohesion. With recent advancements in sophisticated statistical learning approaches (e.g., machine and deep learning algorithms) and the increased use of these approaches in educational text analysis [47], new opportunities have emerged for researchers to investigate and improve their understanding of essay cohesion. Accordingly, in the present study, we attempted to harness machine and deep learning approaches to develop highly accurate models that predict human scores of essay cohesion based on cohesive and other text features. Next, we examined our best-performing predictive models to identify features that may be strongly associated with a human-generated cohesion score.

## 2.2 Machine Learning, Deep Learning and Explainable Artificial Intelligence for Essay Analysis

Researchers have increasingly used machine and deep learning approaches for automated content analysis in student essays over the past years. To date, these approaches have mainly included traditional machine learning methods such as Support Vector Machine (SVM) [21], Random Forest (RF) [43], and XGBoost [17, 35]. Predictive models based on these methods utilized many textual features to evaluate different aspects of essay writing. For example, Hughes et al. [21] computed structural features in a text (e.g., positions of words) to detect causal relations and concepts in explanatory essays. Raković et al. [43] computed rhetorical (e.g., the occurrence of connectives) and content (e.g., the semantic overlap between a sentence and source text) features to predict sentences that transform source information in argumentative essays. Ferreira Mello et al. [17] extracted a comprehensive set of linguistic features (e.g., LIWC [40] and Coh-Metrix [34] indices) to develop machine learning models that predicted the presence of rhetorical categories in college entrance essays.

Recently, researchers have begun utilizing deep learning methods for automated evaluation of student writing. Deep learning algorithms are based on many interconnected neural networks and have been considered an advanced approach in data analytic tasks in different domains [25], including educational research [50]. Whereas predictive models based on deep learning have been shown to generally outperform traditional machine learning models in analyzing student forum posts [47], the use of these models to evaluate student essays remains limited. As a more recent example, Raković et al. [42] developed a deep learning classifier based on the pre-trained language model Legal BERT [7]. The classifier correctly identified more than 86% of rhetorical moves in the legal case note essays written by law students and outperformed traditional machine learning models developed for the same purpose.

The studies mentioned earlier, taken together, have demonstrated the potential of using machine and deep learning methods to automatically and accurately analyze different aspects of student writing based on linguistic features extracted from essays. For this reason, we elected to develop multiple machine and deep learning models in the present study and examine whether those models can accurately predict the essay cohesion score. Thus, our first goal in the current study was to automate the evaluation of essay cohesion.

Researchers often need to explain the prediction output of their predictive models, such as to differentiate between more and less important features in predicting a particular outcome. Machine and deep learning models can be analyzed, and their decisions further "unpacked" and interpreted using explainable artificial intelligence (XAI) methods [49]. Despite the overall importance of explainability in machine and deep learning modelling, learning analytic researchers have yet to use the XAI methods to a larger extent [51].

SHAP (SHapley Additive exPlanations [27]) has been used as one of the XAI methods in learning analytics research to date. For example, Baranyi et al. [2] applied SHAP to their predictive model to understand factors affecting students' graduation probability. For an overview of other XAI methods used in educational research, see [51]. Specifically, SHAP computes an importance value for each feature, representing a conditional contribution of that feature to the model prediction. Lundberg and Lee [27] demonstrated that explanations generated by SHAP were consistent with human explanations. Therefore, XAI methods can provide a detailed overview of the most important features and their effects on outcome variables in different predictive tasks [22]. This opened the possibility of using deep learning in diverse fields, including educational research. With that in mind, we aimed to investigate the potential of SHAP in the present study to reveal and interpret relationships between individual essay features and human ratings of essay cohesion, the second goal of our study.

## 2.3 Research Questions

The first goal of our study was to investigate the viability of using machine and deep learning methods to estimate human ratings of essay cohesion scores. To this end, we extracted an extensive set of linguistic features proposed in prior theoretical and empirical works [17, 18, 24, 38, 40]. The features were extracted from student essays written in Portuguese and English, attempting to add to prior research that appeared to lack a cross-language examination of cohesion. Then, we utilized these features to develop ten machine and deep learning models and compare their performance. More formally, we posed the following research question to guide our investigation at this stage:

> **Research Question 1 (RQ1):**
> *To what extent can machine/deep learning algorithms accurately estimate human-generated scores of essay cohesion?*

The second goal of our study was to understand how individual features affect essay cohesion scores assigned by human raters. To this end, we utilized the SHAP method for explainable AI [27]. More formally, we asked the following research question to guide our investigation at this stage:

> **Research Question 2 (RQ2):**
> *What are the most important features predicting human-generated scores of textual cohesion in essays written in Portuguese and English?*

# 3 METHOD

This study treated the analysis of textual cohesion of essays written in Portuguese and English as a regression problem. To this end, we evaluated two approaches (feature-based and deep learning-based) aiming to develop models capable of estimating a numerical score that reflects the quality of a given essay regarding the cohesion aspect.

## 3.1 Data Description

We obtained two datasets of student essays written in Portuguese and in English [29, 31]. Expert raters evaluated the essays in both datasets, and the scoring rubrics developed for this purpose included different criteria such as content, organization, style, and textual cohesion. Following the goals of the present study, our analyses focused on essay cohesion.

*3.1.1 Portuguese dataset.* The essay writing exam is one of Brazil's most important assessment components of the National High School Exam (ENEM). Millions of prospective university students are required to write a dissertative-argumentative entrance essay on a specific scientific, cultural, political, or social topic. The essay must have a minimum of 8 and a maximum of 30 lines[1]. The Portuguese dataset included the college entrance-level essays from the Essay-BR corpus collected by [29]. This dataset comprised 4,570 essays, divided into 86 topics (prompts), written between December 2015 and April 2020, following the guidelines proposed by the Brazilian National Institute of Educational Studies and Research for the ENEM exam[2]. The essays were extracted from the Vestibular UOL [3] and Educação UOL[4] web portals.

Each essay received an overall score and individual scores for the five competencies assessed in ENEM, including: **(i)** Formal writing (lexical and syntactic aspects); **(ii)** Understanding of the proposed theme; **(iii)** Ability to write a dissertative-argumentative essay; **(iv)** Cohesion of the essay text; and **(v)** Ability to propose an intervention for the problem described in the essay. For each competence, evaluators assigned a score ranging from 0 to 200 in equal intervals of 40. A score of 0 demonstrates a complete absence of knowledge in the competence domain, and a score of 200 demonstrates a complete mastery of a competence. We normalized the scores on the scale to 0 and 1 to facilitate the comparison with essay scores from the English dataset.

*3.1.2 English dataset.* The Automated Student Assessment Prize's (ASAP) dataset is commonly adopted in literature to train and evaluate automated essay scoring systems [31]. This dataset comprises approximately 13,000 essays written in English by 7th-10th grade students following eight prompts. Overall, each essay is approximately 150 to 550 words on average. Only essays from two

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem
[2] https://www.gov.br/inep
[3] https://vestibular.brasilescola.uol.com.br/banco-de-redacoes
[4] https://educacao.uol.com.br/bancoderedacoes /

prompts have individual scores in different aspects, such as content, organization, style, and others. Mathias and Bhattacharyya [31] introduced the ASAP++ dataset, which expands the original ASAP with new annotations to the six prompts that originally only had the general grade score. The essays were scored by expert raters with high competence in English, measured by the scores those raters obtained in high-school English subjects (90% or more) or the TOEFL examination (110 points or more). The original scores attributed by the expert raters range between 0 and 4. However, we normalize the values to a scale of 0 to 1 to facilitate comparison with the Portuguese dataset. This corpus has three essay genres: Argumentative/Persuasive, Source-Dependent Responses, and Narratives.

Among all the essays in the English dataset, we observed that only Source-Dependent Responses were assigned a score for textual cohesion. For this reason, we included the essays of this genre in our analysis. In particular, students who wrote the Source-Dependent Response essays were required to respond to specific questions related to a source text they read, e.g., "Explain the author's position regarding the incident they described in the text.". Each essay was scored following the 4-part scoring rubric: **(i)** Content: relevance of the content; **(ii)** Prompt Adherence: evaluate if the writer answered the question asked; **(iii)** Language: analysis of grammar and spelling mistakes; **(iv)** Narrativity: a measure of the cohesion in the text. We thus gathered and analyzed 7,101 Source-Dependent Response essays. Unfortunately, the paper's authors did not provide details about how the final score was assigned for each essay [31].

*3.1.3 Statistics of the datasets.* In Table 1, we present descriptive statistics about the Portuguese (Essay-BR) and English (ASAP++) datasets, including cohesion scores, the total number of essays, and the mean/standard deviation of sentences and words per essay.

Although the cohesion scores in the datasets adopted for the study were discrete (see Table 1), we decided to explore regression models instead of classifiers. We made this decision based on the literature that treats text cohesion analysis and even essay-scoring systems as regression problems [44]. Therefore, to be consistent with the previous studies, we followed the same approach (e.g., validation process and measures used) [38].

## 3.2 Feature-based Approach

To answer our RQs, we computed a comprehensive set of language-independent features that have typically been adopted in educational text mining and learning analytics research (for an overview, see [47]). For example, some of these features have been previously used to analyze the rhetorical structure of essays [17, 43], automatically score essays [44], analyze online discussions [35], and automatically assess feedback quality [37]. In total, we extracted 183 features describing the Portuguese essays and 208 features describing the English essays. We describe the features in the following subsections.

*3.2.1 Coh-Metrix.* The Coh-Metrix set of linguistic indicators [18, 34] includes several features associated with text cohesion (argument overlap), linguistic complexity (based on syntactic tree structures), text readability (Flesch reading ease), and lexical diversity (type-token ratio). In the present analysis, we adopted the Portuguese version of Coh-Metrix proposed in [5] and the English version proposed in [34].

*3.2.2 LIWC.* In addition to the features provided by Coh-Metrix, we computed the features proposed in the Linguistic Inquiry Word Count (LIWC) dictionary, which includes several indicators related to grammatical, psychological, and social processes that can be inferred from a textual document [52]. We employed the LIWC 2015 [1] and LIWC 2007 [6] versions for English and Portuguese, respectively. The English version of the dictionary adopted here has 93 features, and the Portuguese version has 64 features.

*3.2.3 Lexical Diversity.* The lexical diversity measures indicate how diverse the vocabulary is in each essay. These measures are computed as the ratio between the frequency of different types of words (e.g., nouns, verbs, adjectives, adverbs, and pronouns). Besides, the following three indexes were computed based on the work of Palma and Atkinson [38]: *Hapax Legomena*, *Yule's K*, and *Guiraud's Index*. These indexes complement the previous attributes, as they are computed based on the words and their frequencies without considering their part of speech. In total, 15 lexical diversity features were extracted for Portuguese and English essays.

*3.2.4 Legibility.* The features in this group measure text readability, including indicators that consider aspects such as the number of syllables, word complexity, and the number of characters. For this, the following five attributes were computed [38]: Syllable mean per word, *Flesch Reading Ease*, *Gunning Fog Index*, *Word Variation Index*, and *Automated Readability Index*.

*3.2.5 Sentence Overlap.* Eight sentence overlap indices were extracted to measure referential cohesion in the text. Six features were extracted using the entities grid model that seeks to capture the local cohesion of a text [24]. The intuition of this model is that entities (noun phrases) shared by subsequent sentences contribute to the local cohesion of a text. Thus, the more entities shared between adjacent sentences, the better the local cohesion is. Finally, we calculated the average unigram and cosine similarity between adjacent sentences using the *Term Frequency–Inverse Document Frequency* (TF-IDF) approach.

*3.2.6 Others.* The attributes of this group reflect other, more general aspects of the essays. The following four features were extracted: **(i)** Total number of sentences; **(ii)** Total words classified as *stop words*; **(iii)** Total number of words; and **(iv)** Mean words per sentence.

## 3.3 BERT-based Approach

Recently, learning analytics researchers have begun using pre-trained language models – Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) – to generate features for different tasks, including classification and natural text generation [26]. These models are pre-trained on large collections of textual data. Following the increase in computational power and the emergence of the Transformer architecture [54], the use of pre-trained language models has led to a state-of-the-art performance in many tasks [26].

**Table 1: Descriptive statistics of the Portuguese (Essay-BR) and English (ASAP++) essay datasets.**

| Dataset | Cohesion score | #Essay | Mean Sentences | Mean Words |
|---|---|---|---|---|
| Essay-BR | 0.00 | 134 | 8.80 (3.96) | 216.16 (64.39) |
| | 0.20 | 61 | 9.93 (4.07) | 219.62 (88.75) |
| | 0.40 | 590 | 10.63 (5.67) | 244.32 (90.29) |
| | 0.60 | 2,000 | 11.90 (4.87) | 285.38 (78.78) |
| | 0.80 | 1,241 | 13.37 (4.19) | 325.44 (70.10) |
| | 1.00 | 544 | 13.92 (4.57) | 337.53 (68.21) |
| | *Overall* | *4,570* | *11.26* (4.90) | *294.26 (83.55)* |
| ASAP++ | 0.00 | 1,038 | 2.79 (1.71) | 51.55 (27.28) |
| | 0.25 | 1,843 | 5.47 (2.71) | 97.89 (43.45) |
| | 0.50 | 2,862 | 7.39 (2.92) | 134.61 (47.32) |
| | 0.75 | 1,205 | 9.12 (2.86) | 173.26 (47.44) |
| | 1.00 | 153 | 11.03 (3.61) | 218.73 (57.20) |
| | *Overall* | *7,101* | *6.59 (3.41)* | *121.31 (59.51)* |

Motivated by the recent promising results, we investigated the viability of using the pre-trained language models in the present study, in addition to the feature set we previously extracted following traditional extraction approaches. Specifically, we opted to apply the (BERT) language model [14] to our essay datasets. BERT is a contextual language model that consists of a deep neural network with bidirectional processing. This model can generate embeddings that vary according to the semantic context within which a word occurred, i.e., BERT is sensitive to variations of meaning across texts [14]. Using BERT, we obtained contextual representation of words in our essay datasets, i.e., we obtained features in the form of word embeddings. Then, these features were provided as input to a feed-forward neural network that estimates a cohesion score for an essay.

## 3.4 Model Selection and Evaluation

To address research question RQ1, we trained and evaluated the performance of different traditional regression algorithms using the features described in Section 3.2. The algorithms were trained to estimate the text cohesion scores of essays written in Portuguese and English, i.e., essays in the Essay-BR and ASAP++ datasets, respectively. For this prediction task, we chose algorithms from different mathematical origins, including regression-based approaches, decision trees, classical neural networks, Bayesian models, and ensembles. Ten algorithms were implemented using the scikit-learn library[5], except for the XGB Regressor[6] and CatBoost Regressor[7], which use gradient boosting on decision trees, and LGBM Regressor[8] algorithms. We utilized the default tuning parameters set in the libraries for all the algorithms.

In addition to the traditional regression algorithms, we adopted a BERT-based regression algorithm using a publicly available transformers toolkit[9]. To fine-tune this algorithm, we applied the learning rate scheduler without warm-up, followed by linear decay of the learning rate across the training steps. We used the *BERTimbau Base Cased* [48] and the original *BERT base cased* [14] implementation for Portuguese and English, respectively. The fine-tuning process ran from 1 to 20 epochs evaluating different outcomes. In each epoch, the algorithm run validation every 100 steps at a learning rate of $5 * 10^{-5}$. Then, the tuned BERT model was evaluated on the testing set.

We adopted the evaluation process recommended in the literature [15] to compare the results of the regression models (feature-based and BERT-based). As the number of examples in the evaluated datasets is relatively small, we adopted a 5-*fold Stratified Cross-Validation* as an evaluation methodology for data sampling in combination with the following measures: Pearson correlation, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The RMSE and MAE error measures were computed using the scikit-learn library, and Pearson's correlation was calculated using the SciPy library[10]. Our interpretation of the Pearson correlation follows the guidelines presented in [45], which indicates that: **(i)** 0 value represents no linear relationship; **(ii)** the values +1 and -1 indicate a perfect positive or negative relationship, respectively; **(iii)** values between 0 and 0.3 (0 and -0.3) indicate a weak positive (or negative) relationship; **(iv)** values between 0.3 and 0.7 (-0.3 and -0.7) suggest a moderately positive or negative correlation; and **(v)** values between 0.7 and 1.0 (-0.7 and -1.0) are considered strong positive or negative correlations.

[5]https://scikit-learn.org/
[6]https://github.com/dmlc/xgboost/
[7]https://catboost.ai/
[8]https://github.com/Microsoft/LightGBM/
[9]https://github.com/huggingface/transformers
[10]https://scipy.org/

### 3.5 Feature Importance

To address research question RQ2, we explored the Shapley Additive exPlanations (SHAP) [27] methods for machine learning explainability models to provide insights into the most important features for the predictions made by the models investigated (feature-based and BERT-based). SHAP values can be computed from a group (more generic) or single (focused on a specific outcome) instance and can be explored to analyze the importance and impact of attributes used in traditional machine learning and deep learning models.

More specifically, we performed a three-step process to answer RQ2. The first two steps were related to the feature-based approach combined with the traditional regression models. First, we computed a group analysis of the SHAP values. Thus, we applied SHAP for each test set generated during the 5-fold Stratified Cross Validation (see Section 3.4 for details), and then we obtained the mean of the feature's importance for each interaction. Second, we analyzed the most relevant features for two essay predictions, one from each dataset (Essay-BR and ASAP++). The second investigation could provide specific information to support student feedback as the feature importance is extracted from the essay of a specific student.

The last step in the feature importance analysis was the application of SHAP for the BERT-based model. As BERT is based on words, not general indicators, we only computed the word importance focused on a specific essay (single prediction). The same essays selected in the feature-based approach were used in the approach using BERT to allow a comparison between the two approaches.

## 4 RESULTS

### 4.1 RQ1: To what extent can machine/deep learning algorithms accurately estimate human-generated scores of essay cohesion?

Table 2 presents RMSE, MAE, and Pearson correlation scores obtained by conventional machine learning algorithms considered in our study. The CatBoost Regressor was the best-performing model on the ASAP++ dataset, achieving an RMSE value of less than 0.16. In the Essay-BR dataset, CatBoost Regressor and Bayesian Ridge presented the best results with an RMSE of nearly 0.18. We selected the CatBoost Regressor as the best model identified in the performed experiments. The average cohesion scores in the English and Portuguese essay datasets were 0.415 and 0.653, respectively. The estimated error percentage for the best-performing model on these datasets was 38.07% and 27.41%, respectively.

The Pearson correlation of approximately 0.77 and 0.53 obtained by CatBoost Regressor on the ASAP++ and Essay-BR datasets indicate moderate to strong positive correlations between computer- and human-generated cohesion scores for essays written in English and Portuguese. Overall, the results suggest that the cohesion scores estimated by the best-performing traditional machine learning algorithm in our study tend to provide a considerably accurate estimation of human-generated cohesion scores in both datasets.

Next, we evaluated the performance of the BERT-based deep learning model in estimating the essay cohesion score. Table 3 shows the performance metrics (RMSE, MAE, and Person correlation) that this model achieved across multiple epochs. On the Essay-BR dataset, the BERT-based model with 20 epochs achieved the best performance according to all the metrics. On the ASAP++ dataset, the BERT-based model with one epoch achieved the best performance according to all the metrics except for MAE, as the best MAE value was reached after 20 epochs. We note that the BERT-based model outperformed all the conventional machine learning models we analyzed in this study.

### 4.2 RQ2: What are the most important features predicting human-generated scores of textual cohesion in essays written in Portuguese and English?

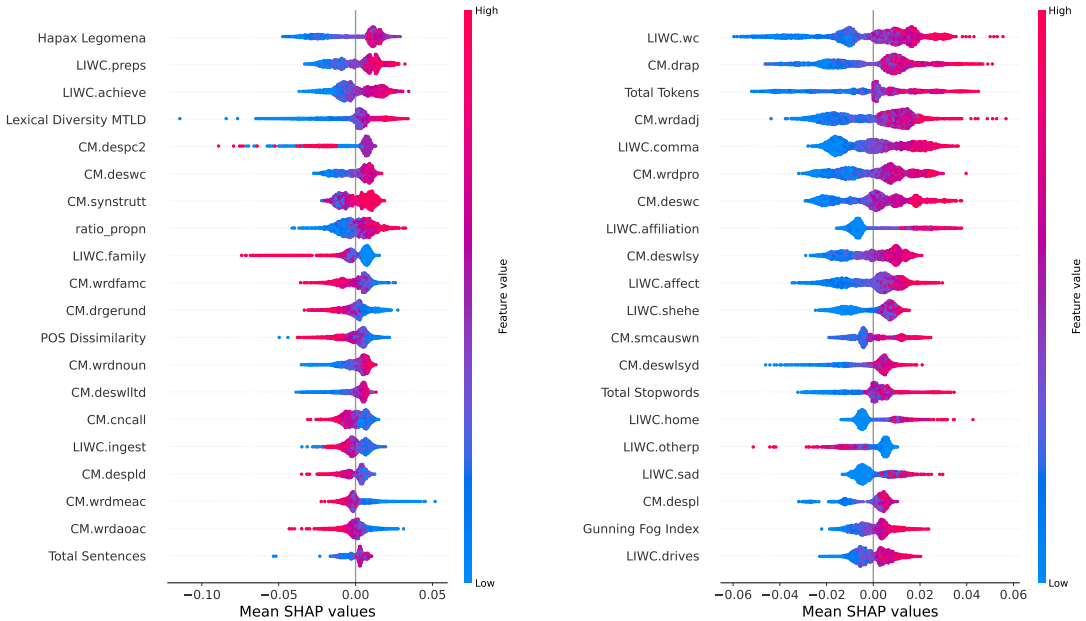
To answer research question RQ2, we analyzed the most relevant features in both Essay-BR and ASAP++ datasets using the CatBoost Regressor, our study's best-performing conventional machine learning model. Following the process described in Section 3.5, we performed a two-step analysis for the traditional regression models. Figure 1 presents an overview of the mean SHAP values of the top 15 features based on their importance for the CatBoost Regressor predictions for each test set generated during the 5-fold stratified cross-validation on Essay-BR and ASAP++, respectively. Figure 1a shows that *Hapax Legomena* – i.e., the number of lemmatized words that occur only once in the essay – was the most relevant feature for the Essay-BR database. In particular, the graph indicates that essays with a high *Hapax Legomena* value were likely to receive high cohesion scores. The effect of the *CM.drgerund* feature was the opposite, i.e., essays that contained many verbs in the gerund tended to receive lower cohesion scores.

Figure 1b shows that *LIWC.wc* – the feature representing the word count computed by the LIWC – was identified as the most relevant feature for the essays in the ASAP++ dataset. Essays with more words hence tended to have high cohesion scores. Except for the *Total Tokens*, *Total Stopwords* and *Gunning Fog Index* features, all other features identified as important were extracted from LIWC and Coh-Metrix groups, which confirms the importance of these groups of features in understanding student essays, including their cohesion quality.

In the second step of this analysis, we assessed the most important features of individual essays. Figure 2a presents the SHAP values to explain the impact (positive or negative) of the features on a single cohesion score prediction in an essay extract from Essay-BR. The color in the figure indicates whether the final score increased (red) or decreased (blue). In the selected Essay-BR example, the final prediction was 0.699 (the actual score is 0.6). We also note that the effects of some features in the individual analysis did not necessarily match the effects of those same features in the general analysis (Figure 1a). However, it is possible to see that the *Hapax Legomena* (i.e., the most important feature in the general analysis) was also considered influential with a positive impact on estimating the final score of this specific essay in the individual analysis. On the other hand, the *LIWC.ingest* (i.e., words related to the ingestion of food) was selected as the most important feature with a negative impact in this case. Regarding the ASAP++ example (Figure 2b), the model estimated the cohesion score for that essay to be 0.425. The real value, however, was 0.25. The *CM.wrdpro* (i.e., the incidence of personal pronouns per 1000 words.) was the feature

**Table 2: Performance of the traditional machine learning models in estimating essay cohesion score on the Essay-BR and ASAP++ datasets.**

| Algorithms | Essay-BR | | | ASAP++ | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | P | RMSE | MAE | P |
| AdaBoost | 0.194 | 0.151 | 0.475 | 0.181 | 0.151 | 0.702 |
| Bayesian Ridge | **0.179** | **0.134** | **0.536** | 0.161 | 0.128 | 0.760 |
| CatBoost Regressor | **0.179** | **0.134** | **0.536** | **0.158** | **0.126** | **0.770** |
| Extremely Randomized Trees | 0.180 | 0.135 | 0.531 | 0.161 | 0.128 | 0.760 |
| Gradient Boosting | 0.185 | 0.139 | 0.508 | 0.163 | 0.130 | 0.754 |
| LGBM Regressor | 0.181 | 0.137 | 0.523 | 0.160 | 0.127 | 0.765 |
| Linear Regression | 0.180 | 0.136 | 0.533 | 0.161 | 0.128 | 0.760 |
| MLP | 0.192 | 0.146 | 0.491 | 0.182 | 0.144 | 0.712 |
| Random Forest | 0.180 | 0.136 | 0.526 | 0.162 | 0.129 | 0.758 |
| XGB Regressor | 0.192 | 0.145 | 0.464 | 0.170 | 0.134 | 0.732 |

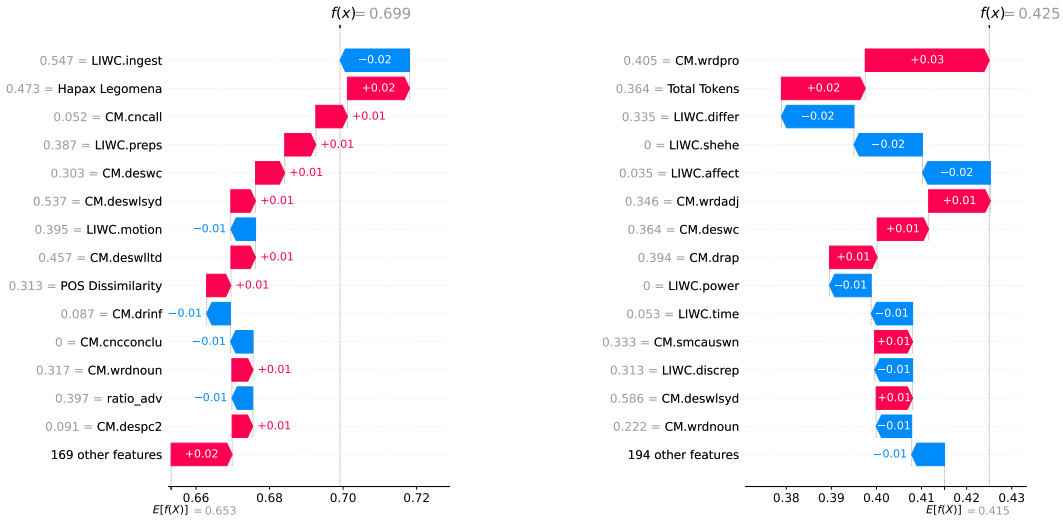



**(a) Essay-BR** **(b) ASAP++**

**Figure 1: Mean SHAP values of the group feature analysis from (a) Essay-BR and (b) ASAP++ based on CatBoost Regressor predictions for each test set generated during the 5-fold stratified cross-validation.**

that had the most significant positive impact on estimating the final grade, whereas three features extracted from the LIWC dictionary (*LIWC.shehe*, *LIWC.differ* and *LIWC.affec*) showed the greatest negative influence.

Finally, the last step of the features importance analysis was assessing the BERT outcome. Figure 3 shows the real and predicted values for the text cohesion score of each essay. The words that positively impacted the final prediction are highlighted in green, whereas those that negatively impacted the final prediction are highlighted in red. The figures show excerpts of the essays written in Portuguese and English. In contrast to the previous analysis (Figure 2), Figure 3 is focused on assessing the importance of the words in the text to predict textual cohesion.




**(a) Essay-BR** **(b) ASAP++**

**Figure 2: Features impact on the prediction of a single essay based on SHAP values using CatBoost Regressor. The red and blue colours represent a positive and negative influence in the final score, respectively.**

**Legend:** Negative Neutral Positive

| True Grade | Predicted Grade | Word Importance |
|---|---|---|
| 0.6 | 0.70 | [CLS] Em 2018 , o Brasil encontrou - se em uma eleição presidencial extrem ##ista , em que as ideologia ##s dos candidatos foram totalmente contr ##árias entre si . Neste contexto , a discussão sobre a " onda conservadora " no país foi leva ##ntada , juntamente com as questões morais da massa populacional , que , vis ##lu ##mbra ##ndo a decisão dentre os extremos , tomou partido em questões como , a pena de morte , prisão perpé ##tua , casamento homossexual e tantas outras polêmica ##s , que anteriormente não eram leva ##ntadas com tanta frequência . Segundo [SEP] |

**(a) Essay-BR**

**Legend:** Negative Neutral Positive

| True Grade | Predicted Grade | Word Importance |
|---|---|---|
| 0.25 | 0.43 | [CLS] The features of the setting affect the cyclist in two way . This ways are p ##ys ##ical and mental . The p ##ys ##ical ways that features affect cyclist is the terrain . If the roads are long and wired ##y , this could wear the cyclist out . They could also become physically tired if the in ##cline of the hill is to steep . When its you against gravity , it can be e ##gau ##sting . The temperature can also be a c ##rus ##al factor in a cyclist p ##ys ##ical ability . Nobody want to feel like their in a sa ##una when they are bike riding . The cyclist and [SEP] |

**(b) ASAP++**

**Figure 3: Features importance based on SHAP values using BERT. The green and red colors represent a positive and negative influence in the final score, respectively.**

## 5 DISCUSSION

Cohesion is considered one of the most important characteristics of well-written essays [11], as it ensures that the ideas in the text are connected into a coherent whole. With this in mind, researchers developed several computational approaches, e.g., [12, 18], as an attempt to automate the evaluation of cohesion in student essays and

**Table 3: Performance of the BERT-based deep learning model in predicting essay cohesion score on Essay-BR and ASAP++ datasets.**

| Epochs | Essay-BR | | | ASAP++ | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | P | RMSE | MAE | P |
| 1 | 0.169 | 0.122 | 0.641 | **0.149** | 0.119 | **0.804** |
| 5 | 0.165 | 0.119 | 0.658 | 0.156 | 0.123 | 0.794 |
| 10 | 0.167 | 0.121 | 0.656 | 0.156 | 0.122 | 0.797 |
| 15 | 0.165 | 0.118 | 0.666 | 0.159 | 0.121 | 0.791 |
| 20 | **0.164** | **0.117** | **0.670** | 0.160 | **0.115** | 0.7817 |

provide appropriate support to students. However, in multiple empirical studies, researchers have documented inconsistent relationships between computationally extracted text features and human ratings of cohesion. To advance understanding of the association between text features and cohesion, we extracted a comprehensive set of theoretically supported linguistic features from essays written in Portuguese and English, developed state-of-the-art predictive models that estimate essay coherence, and examined features in these models using methods in explainable artificial intelligence.

The deep learning model based on BERT outperformed the conventional machine learning models in estimating the cohesion score for the essays written in both languages. We expected this result, as prior literature suggests that deep learning models generally tend to outperform traditional machine learning models in analyzing educational texts [25, 50]. On the other hand, CatBoost Regressor, our study's best-performing conventional machine learning model, demonstrated to estimate the cohesion scores with a moderate positive Pearson correlation in the essays written in Portuguese and English. Given that (1) the BERT model relies upon content-dependent features, i.e., features driven by the essay text and not by theory, unlike e.g., Coh-Metrix and LIWC features, and (2) the conventional machine learning regressor in our study demonstrated an attractive prediction accuracy, we posit the conventional model may be a more viable solution to automate coherence scoring and provide writing evaluation to students based on text-independent and theoretically supported features it was trained on.

Among the essays written in Portuguese, human raters tended to assign higher cohesion scores to dissertative-argumentative essays that were lexically more diverse. This finding challenges commonly adopted views on text coherence. More specifically, as lexical repetition (e.g., reusing the same words and phrases throughout the text) is theorized to contribute to text cohesion [53], it may be expected that the increase of lexical diversity of a text may be related to decrease of text cohesion. The unexpected results from our analysis may indicate that human raters who scored Brazilian college entrance essays were experts in the essay genre and topics and, as such, they did not rely on rhetorical signals for cohesion (e.g., rhetorical connectives) to mentally create a coherent picture of an essay argument, which, in turn, aligns with the position in [32].

Among the essays written in English, human raters tended to provide higher cohesion scores to longer submissions, as indicated by two of the most important features in this dataset, i.e., the word count measured by LIWC and the total number of tokens in the essay. Given that in this writing task, students were required to describe the situation from an assigned source article, this finding may indicate that students who wrote longer texts were more successful in coherently describing the situation than their counterparts. In other words, writing longer texts may provide students with more opportunities to develop cohesive ties within those texts [20]. The higher presence of adjectives and adverbial phrases was also related to higher scores for cohesion. As these types of words additionally describe properties of noun and verb phrases, respectively [18], we posit those may have contributed to creating coherent story source-dependent essays. This hypothesis should be investigated in future research.

## 6 LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

We acknowledge the following limitations of the study. First, although we have explored a wide range of traditional machine learning algorithms and a deep learning language model, we have not explored methods related to parameter tuning and data balancing. These methods could improve the outcome of the algorithms. In future works, we intend to explore methods proposed in previous works [3, 36]. Second, the data used in the evaluation comprised different characteristics, language (Portuguese and English), level of education (high school and secondary/mid-school students), an average of words per essay (294.26 and 121.31 for the Portuguese and English data, respectively), topics and tasks. It indicates the potential of the proposed approach in terms of generalizability. However, we have not performed a specific analysis to evaluate this issue. We intend, as future work, to perform a detailed assessment to measure the generalizability of our approach. Finally, this study did not intend to analyze the practical application of the proposed method. It would include the development of a learning analytics tool and assessing instructors' and students' satisfaction based on the output of our method. We intend to develop such a tool for future line research.

## REFERENCES

[1] Pedro Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

[2] Máté Baranyi, Marcell Nagy, and Roland Molontay. 2020. Interpretable deep learning for university dropout prediction. In *Proceedings of the 21st annual conference on information technology education*. 13–19.

[3] Gian Barbosa, Raissa Camelo, Anderson Pinheiro Cavalcanti, Péricles Miranda, Rafael Ferreira Mello, Vitomir Kovanović, and Dragan Gašević. 2020. Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 605–614.

[4] Nadia Behizadeh and Myoung Eun Pang. 2016. Awaiting a new wave: The status of state writing assessment in the United States. *Assessing Writing* 29 (2016), 25–41.

[5] Raissa Camelo, Samuel Justino, and Rafael Ferreira Leite de Mello. 2020. Coh-Metrix PT-BR: Uma API web de análise textual para a educação. In *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*. SBC, 179–186.

[6] Flavio Carvalho, Rafael Guimarães Rodrigues, Gabriel Santos, Pedro Cruz, Lilian Ferrari, and Gustavo Paiva Guedes. 2019. Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*. SBC, 24–34.

[7] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559* (2020).

[8] Elena Cotos, Sarah Huffman, and Stephanie Link. 2020. Understanding Graduate Writers' Interaction with and Impact of the Research Writing Tutor during Revision. *Journal of Writing Research* 12, 1 (2020).

[9] Scott Crossley and Danielle McNamara. 2010. Cohesion, coherence, and expert evaluations of writing proficiency. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 32.

[10] Scott Crossley and Danielle McNamara. 2011. Text coherence and judgments of essay quality: Models of quality and coherence. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.

[11] Scott A Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research* 11, 3 (2020), 415–443.

[12] Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods* 48, 4 (2016), 1227–1237.

[13] Scott A Crossley and Danielle S McNamara. 2016. Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research* 7, 3 (2016), 351–370.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] Manuel Fernández-Delgado, Manisha Sanjay Sirsat, Eva Cernadas, Sadi Alawadi, Senén Barro, and Manuel Febrero-Bande. 2019. An extensive experimental survey of regression methods. *Neural Networks* 111 (2019), 11–34.

[16] Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 6 (2019), e1332.

[17] Rafael Ferreira Mello, Giuseppe Fiorentino, Hilário Oliveira, Péricles Miranda, Mladen Rakovic, and Dragan Gasevic. 2022. Towards Automated Content Analysis of Rhetorical Structure of Written Essays Using Sequential Content-Independent Features in Portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. Association for Computing Machinery, New York, NY, USA, 404–414. https://doi.org/10.1145/3506860.3506977

[18] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 2 (2004), 193–202.

[19] Daniela Faria Grama. 2022. Elementos coesivos do português brasileiro em córpus de redações nos moldes do Enem: um estudo para a elaboração da CoTex. (2022).

[20] Richard H Haswell. 1988. Critique: Length of text and the measurement of cohesion. *Research in the Teaching of English* (1988), 428–433.

[21] Simon Hughes, Peter Hastings, Mary Anne Britt, Patricia Wallace, and Dylan Blaum. 2015. Machine learning for holistic evaluation of scientific essays. In *International Conference on Artificial Intelligence in Education*. Springer, 165–175.

[22] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence* 3 (2022), 100074.

[23] Simon Knight, Antonette Shibani, Sophie Abel, Andrew Gibson, Philippa Ryan, Nicole Sutton, Raechel Wight, Cherie Lucas, Agnes Sandor, Kirsty Kitto, et al. 2020. AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research* 12, 1 (2020), 141–186.

[24] Mirella Lapata and Regina Barzilay. 2005. Automatic Evaluation of Text Coherence: Models and Representations. In *IJCAI*. 1085–1090.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[26] Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji rong Wen. 2021. Pretrained Language Models for Text Generation: A Survey. In *IJCAI*.

[27] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[28] Charles A MacArthur, Amanda Jennings, and Zoi A Philippakos. 2019. Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing* 32, 6 (2019), 1553–1574.

[29] Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2021. Essay-BR: a Brazilian Corpus of Essays. In *Anais do III Dataset Showcase Workshop* (Rio de Janeiro). SBC, Porto Alegre, RS, Brasil, 53–64. https://doi.org/10.5753/dsw.2021.17414

[30] Mar Mateos, Isabel Solé, Elena Martín, Isabel Cuevas, Mariana Miras, and Nuria Castells. 2014. Writing a synthesis from multiple sources as a learning activity. In *Writing as a learning activity*. Brill, 169–190.

[31] Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan). 1169–1173.

[32] Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication* 27, 1 (2010), 57–86.

[33] Danielle S McNamara, Scott A Crossley, and Rod Roscoe. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods* 45, 2 (2013), 499–515.

[34] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

[35] Rafael Ferreira Mello, Giuseppe Fiorentino, Péricles Miranda, Hilário Oliveira, Mladen Raković, and Dragan Gašević. 2021. Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays. In *International Conference on Artificial Intelligence in Education*. Springer, 162–167.

[36] Péricles BC Miranda, Rafael Ferreira Mello, André CA Nascimento, and Tapas Si. 2022. Multi-Objective Optimization of Sampling Algorithms Pipeline for Unbalanced Problems. In *2022 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1–8.

[37] Ikenna Osakwe, Guanliang Chen, Alex Whitelock-Wainwright, Dragan Gašević, Anderson Pinheiro Cavalcanti, and Rafael Ferreira Mello. 2022. Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence* 3 (2022), 100059.

[38] Diego Palma and John Atkinson. 2018. Coherence-based automatic essay assessment. *IEEE Intelligent Systems* 33, 5 (2018), 26–36.

[39] Yo-Han Park, Yong-Seok Choi, Cheon-Young Park, and Kong-Joo Lee. 2022. EssayGAN: Essay Data Augmentation Based on Generative Adversarial Networks for Automated Essay Scoring. *Applied Sciences* 12, 12 (2022), 5803.

[40] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[41] Dolores Perin and Mark Lauterbach. 2018. Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education* 28, 1 (2018), 56–78.

[42] Mladen Raković, Lele Sha, Gerry Nagtzaam, Nick Young, Patrick Stratmann, Dragan Gašević, and Guanliang Chen. 2022. Towards the Automated Evaluation of Legal Casenote Essays. In *International Conference on Artificial Intelligence in Education*. Springer, 167–179.

[43] Mladen Raković, Philip H Winne, Zahia Marzouk, and Daniel Chang. 2021. Automatic identification of knowledge-transforming content in argument essays developed from multiple sources. *Journal of Computer Assisted Learning* (2021).

[44] Dadi Ramesh and Suresh Kumar Sanampudi. 2021. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review* (2021), 1–33.

[45] Bruce Ratner. 2009. The correlation coefficient: Its values range between+ 1/- 1, or do they? *Journal of targeting, measurement and analysis for marketing* 17, 2 (2009), 139–142.

[46] Fátima Rodrigues and Paulo Oliveira. 2014. A system for formative assessment and monitoring of students' progress. *Computers & Education* 76 (2014), 30–41.

[47] Lele Sha, Mladen Rakovic, Yuheng Li, Alexander Whitelock-Wainwright, David Carroll, Dragan Gaševic, and Guanliang Chen. 2021. Which Hammer Should I Use? A Systematic Evaluation of Approaches for Classifying Educational Forum Posts. *International Educational Data Mining Society* (2021).

[48] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems*. Springer, 403–417.

[49] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2239–2250.

[50] Zhongtian Sun, Anoushka Harit, Jialin Yu, Alexandra I Cristea, and Lei Shi. 2021. A brief survey of deep learning approaches for learning analytics on MOOCs. In *International Conference on Intelligent Tutoring Systems*. Springer, 28–37.

[51] Vinitra Swamy, Bahar Radmehr, Natasa Krco, Mirko Marras, and Tanja Käser. 2022. Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. *arXiv preprint arXiv:2207.00551* (2022).

[52] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. https://doi.org/10.1177/0261927X09351676

[53] Andrea Tyler. 1994. The role of repetition in perceptions of discourse coherence. *Journal of Pragmatics* 21, 6 (1994), 671–688.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008.